Evolution of a family of virulent factors in pathogenic bacteria

Project by: Anagha Kadam
Advisor: Dr. N. Luisa Hiller

Introduction

The emergence of multi-drug resistance in invasive strains of pathogenic bacteria is surfacing as a major challenge in bacterial pathology. Advancements in sequencing technologies have begun to reveal the existence of in-strain genetic heterogeneity within bacterial populations leading to the idea of distributed genome[1,2]. This hypothesis states that, outside of the common core genes, there are distributed genes that are not shared among all the strains in a microbial population[3]. The core and the distributed genes together constitute a microbial supragenome or pangenome[4]. This genetic heterogeneity may have taxonomic and pathogenic implications, giving opportunities to evade treatment strategies. The virulence of a microbe also depends on the virulent and commensal microbiota occupying the same niche and possible genetic exchanges may contribute to the pathogenic status of the microbe[5]. This makes for an interesting problem to track the evolution of virulence factors in related bacterial species and strains.

Haemophilus influenzae is a Gram negative coccobacillus and an opportunistic pathogen affecting the upper and lower respiratory tracts. It is known to cause severe paediatric illnesses like meningitis, otitis media (OM), pneumonia and chronic obstructive pulmonary disease (COPD)[6]. Recent studies indicate the emergence of antibiotic-resistant and invasive nontypeable *Haemophilus influenze* (NTHi) strains that present a major public health challenge[7]. Many pathogenic bacteria species code for a family of proteins, called SLR (Sel like repeat) characterized by multiple 36-residue repeats. Presence of the SLR family has been shown to be associated with virulence in *Legionella pneumophilla*, *Helicobacter pylori*, and most recently NTHi[8,9,10,11]. In NTHi, the SLR family comprises of seven subfamilies. Individual strains of NTHi contain variable numbers of SLR subfamilies (0-4 per strain), diverse numbers of proteins from each subfamily (up to four proteins from the same family per strain), and single SLR proteins contain distinct numbers of repeats (2-10 repeats per protein)[8]. All the variables generate high diversity among NTHi strains. Recent data from SLR analysis in NTHi indicates that the presence of subfamily 1 co-relates with virulence. Further, subfamilies 2 and 3 are only observed in combination with subfamily 1, and when present share the same locus. The other four subfamilies (4, 5, 6 and 7) show a negative co-relation with virulence and are less frequently observed. These differences suggest that the different families are associated with different functions.

BLAST searches querying the NTHi SLR proteins against the non-redundant database have revealed that some subfamilies have highly related sequences in other species, such that there may be higher similarity across species than among strains of the same species. From this observation, we hypothesize that the evolution of SLR families in bacterial pathogens proceeded through horizontal gene transfer events leading to mixing of these genes across bacterial species. If the genes were not transferred across the species (or if transferred

occurred among ancestors and before the speciation event), one would expect all SLR genes from a single species to cluster together. Additionally, the gene tree and the species tree would overlap, indicating parallel evolution of the gene and species divergence. If, however, the intermixing of SLR genes is a result of horizontal gene transfer as we hypothesize, one would see ectopic insertions of out-species SLR sequences into other species in a gene tree, indicating closer relations that have not arisen from common species ancestry.

Based on recently generated unpublished data on the SLRs in NTHi[8], the known relationships between pathogens, and performing BLAST searches, the subjects of this study were SLR sequences from representative strains of some pathogenic bacterial species.

Methods

Procuring the data set

To generate a robust data set, BLAST searches were performed against the non-redundant database using one representative sequence from each of the seven NTHi subfamilies as query. The species with highest similarity to the NTHi SLRs were selected, leading to five species of pathogenic bacteria, with multiple strains for some, resulting in ninety (90) sequences (Table 1). Twenty strains of *H. influenzae* from the previous study were used. These twenty NTHi strains show combination of distributions of SLR sequences from within the seven subfamilies. Further, since the SLR system of genes has been previously characterized in *H. pylori* and *L. pneumophila*[9,10,11], these were also included in the data set.

Multiple Sequence Alignment

Sequence alignment of repeats is a challenging task for the commonly used alignment tools. We sought out to align sequences using T-Coffee in the Jalview package with special considerations on anchoring our multiple alignment on some key residues to give the best alignment. First, a couple of very long sequences that were disturbing the alignment were eliminated from the analysis (but other genes from these strains were still included in the analysis)(Fig.1). Next, the C-terminal of the SLR protein sequences was anchored on two highly conserved cysteine residues. The alignment of the individual repeats were anchored on glycine and alanine residues that confer important structural properties to SLR gene products and are therefore conserved. Furthermore, some highly conserved amino acid stretches with the SLR repeats (such as the RKKAA in NTHi) were also used to confirm the alignment of the repeats. The alignment output was subjected to manual trimming to eliminate uninformative gaps in the sequences (Fig.2). In general, a column with more than 50% gaps was deleted with the exception of high occurrence of SLR relevant stretches in the remaining sequences (that is, SLR repeats from the sequences with high numbers of repeats).

Model Generator:

TopaLi was used to select the best model for tree construction for this data. Based on the value of both, AIC and BIC information criterion, the best model for amino acid equilibrium frequencies for this data was found to be WAG+G (Fig.3).

Phylogenetic tree construction:

PhyML software from the Seaview interface was used for tree construction. WAG model with a fixed value of across site variations and 100 bootstraps was selected. The obtained gene tree was viewed and annotated in Figtree for better interpretation (Fig.4).

Preliminary estimation of gene transfers:

For an introductory understanding of horizontal transfers in SLR genes across species, Notung 2.7 beta version was used. The costs assigned for Notung analysis were 2.0 for losses, 3.0 for duplication and 4.0 for transfers. The appropriate species tree was obtained from NCBI Taxonomy database (Fig. 5). Note however, that this tree includes only one strain representative per species. Relevant changes were made in the nomenclatures and annotations of the gene and species taxa to allow for reconciliation of the trees. Specifically, SLR sequences from different NTHi strains were grouped together into a single species entry. Yet, the seven NTHi subfamilies were differentiated. Thus, the results will not allow for any conclusion regarding NTHi strain variation. It is realized that this assumption does not lead to an accurate and conclusive idea of genetic events, the tree being biased to interpret duplications, since all strains are grouped into one. Nevertheless, this allows an observation of transfer events across species. If these are observed, they provide some leads and corroborate the need for more extensive and sincere analysis.

Results

SLR gene products exist in exclusive, unmixed gene clusters

From the gene tree, it was noted that the SLR subfamilies 1 through 7 of NTHi exist as separate groups and there is low similarity between the amino acid sequences between individual subfamilies.

Existence of greater similarity among SLRs across species than within species

Haemophilus haemolyticus and *Haemophilus influenzae* are closely related species sharing similar cell morphologies, biochemical characteristics and same niche in respiratory tract[12]. The gene tree from this study indicates that SLR-like sequences of *H. haemolyticus* strains have close similarity to specific subfamilies of NTHi SLRs. Three sequences from strains

M21621 and M19501 were found to be closely related to NTHi subfamily 1 of SLR. Likewise, other NTHi subfamilies were also closely related to *H. haemolyticus* sequences via a common parent- one in NTHi subfamily 2, one in subfamily 6 and one in subfamily 7. It is interesting to note that NTHi SLRs share more similarity with SLRs in other species than they share within subfamilies in NTHi. This hints towards genetic events conferring gene transfers probably after speciation events.

In addition to similarities between SLRs in closely related species, the tree infers close relations between SLRs across more distanced species- *Kingella oralis* and *Legionella pneumophila* that to family Nesseriaceae and Gammaprotobacteria respectively. SLR sequences of Helicobacter pylori appear to be clustered separately and do not show intermixing with other species. Some sequences of *Neisseria* remain separate, while some show links to *Kingella denitrificans*, another member of Nesseriaceae family.

Preliminary estimation of SLR gene transfers

Notung 2.7 was used to infer transfers in SLRs. Importantly, SLR genes from different NTHi strains were all treated as a single species entry. Notung inferred duplications at many nodes, and many of these are likely an artefact of grouping together sequences from multiple strains. This simplification should still allow us to draw conclusion about cross-species transfers, which was the main focus of this project. The Notung parameters were set by allocating costs of 2.0, 3.0 and 4.0 to losses, duplications and transfers respectively. Corroborating the previous view, the reconciled tree inferred transfer from NTHi subfamilies into *Haemophilus haemolyticus*. Also noted were the larger distance transfers between *Kingella* and *Legionella* (Fig.6). Since an unrooted tree was used in the analysis, the outcome of Notung could have been altered and some transfers depicted “into” NTHi subfamilies could be an artefact.

Discussion

The occurrence of different topologies of gene tree and the species tree suggest evolution of SLR genes independent of divergence of species. Taken together, the observations from gene tree and the reconciled tree suggest that events of gene transfer between our test species occurred after speciation event. Horizontal gene transfers may lead to transfer of genes that confer selective advantage of some kind to the recipient. Many commensal or less virulent bacteria co-exist with invasive pathogens and are utilized as genetic reservoirs to modulate virulence of pathogenic neighbours. It is known that *Haemophilus haemolyticus* is a less invasive pathogen than *Haemophilus influenze*[13]; this may underlie the less extensive presence of SLR virulent gene products in *H. haemolyticus*. Some degree of SLR genes transferred to *H. haemolyticus* may have equipped it with a partial virulence and

survival potential to ride out host environment shared by NTHi and continue to act as a genetic reservoir. In other context, gene transfer mechanisms contribute to random genetic variability that in advantageous for the pathogen to evade the host immune system.

Future directions

In this study, an important consideration is that *H. influenzae* (24 strains) have been classified based on SLR subfamilies. We were not interested in within-strain differences in *H. influenzae* for this particular exercise. That question would form another layer of scientific problem that would be addressed in the future work. Appropriate reconciliation of trees will be undertaken after producing a higher resolution species tree that reads out differences between strains within a species.

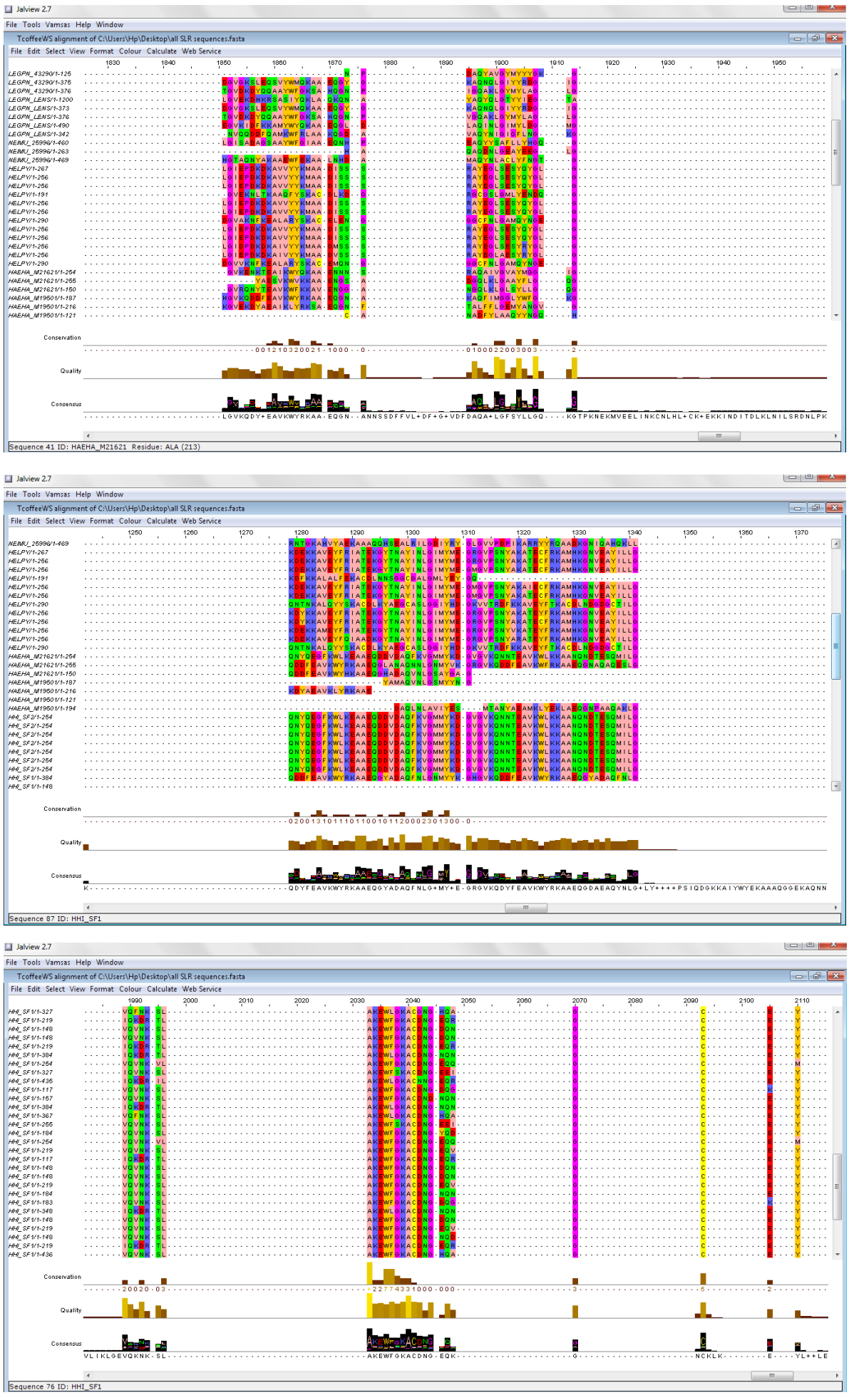


Figure 1. Multiple Sequence Alignment. Three sections of T-Coffee multiple sequence alignment showing that conservation was observed in blocks of matching sequences, which is expected when aligning repeats with variable units in each sequence.

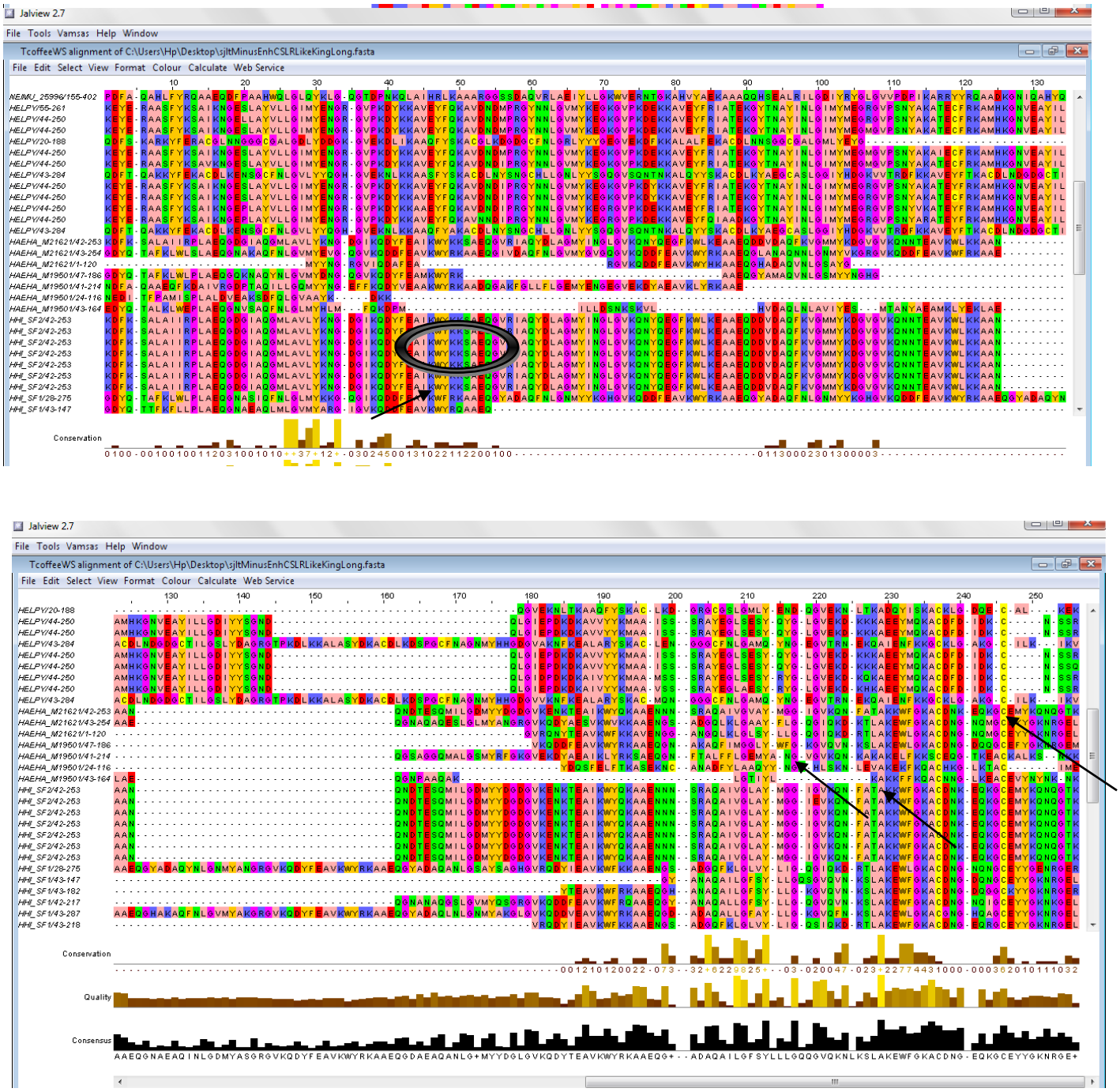


Figure 2. Trimmed SLR alignment. The alignment was anchored at conserved cysteines (C), glycines (G) and alanines (A) (shown with arrows) and the KKAA repeats preceded by conserved tyrosine (Y) and tryptophan (W) residues.

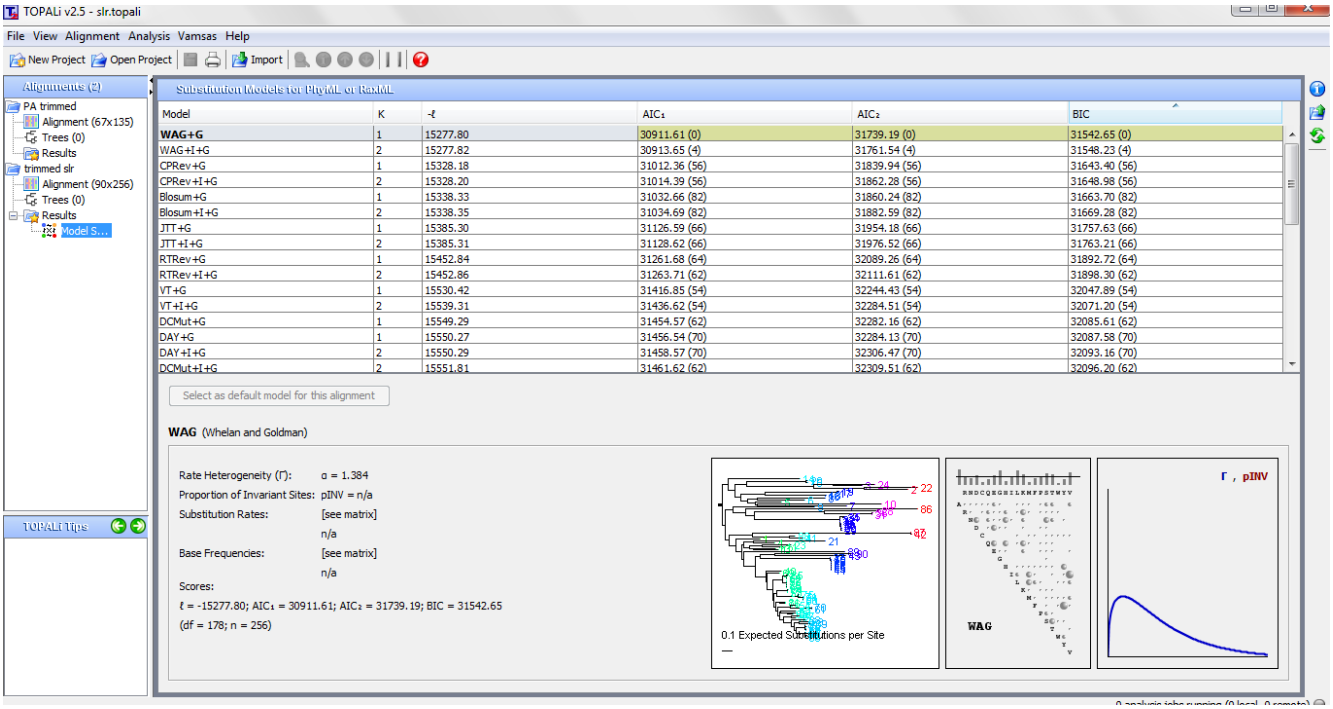


Figure 3. Model Selection. TopaLi model generator output suggesting WAG to be the best model for phylgeny reconstruction of the given data

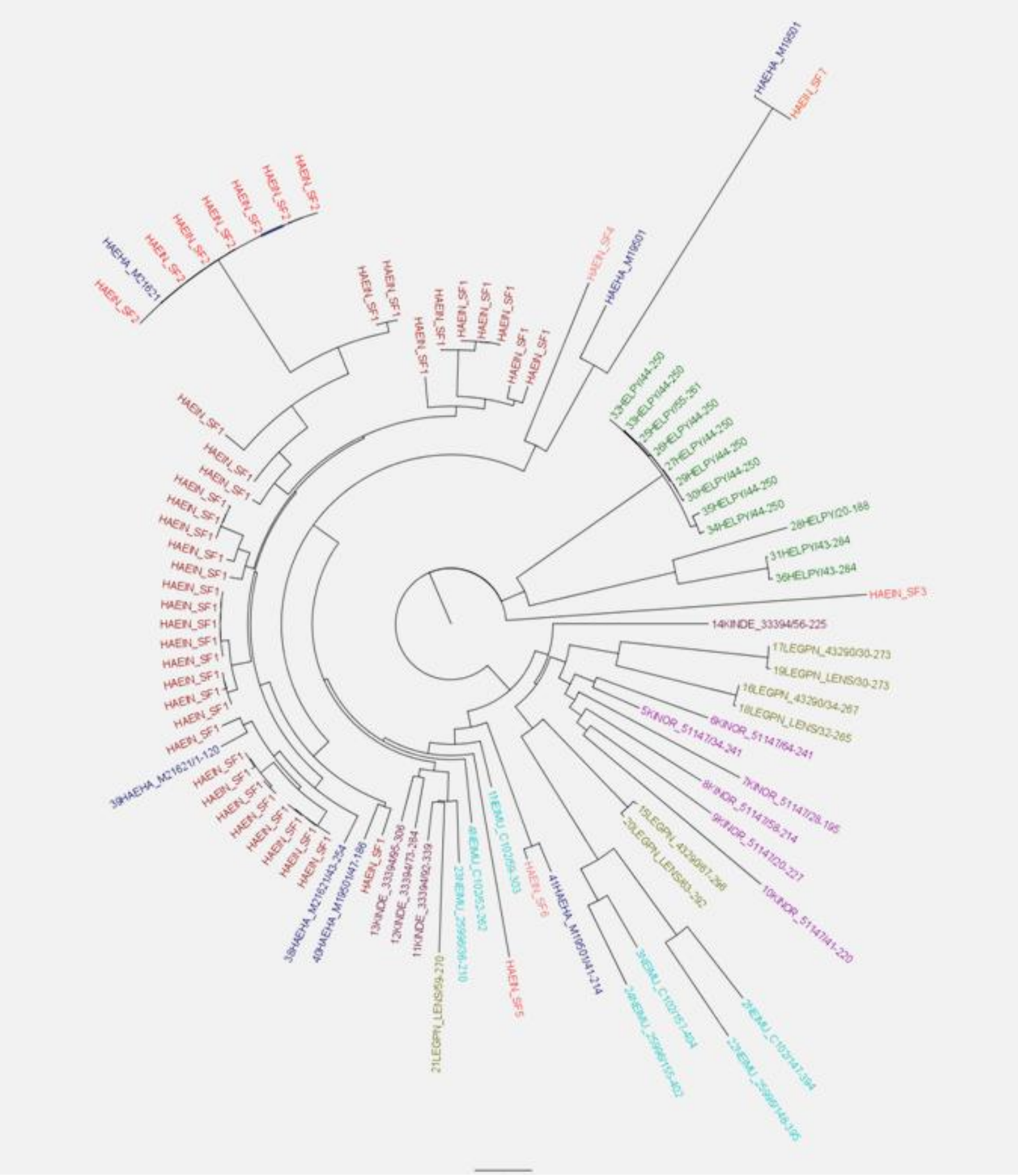
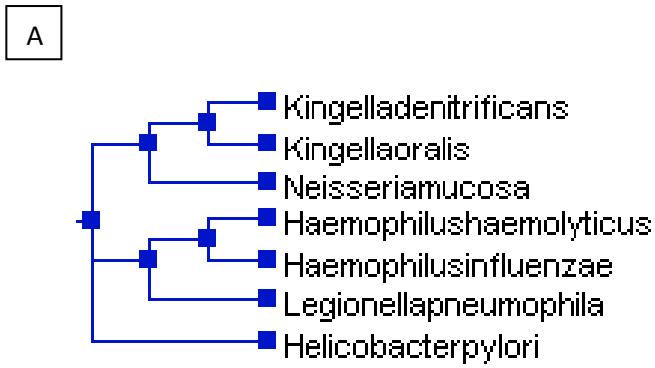


Figure 4. Gene tree as obtained from PhyML (Seaview) and viewed in Figtree. Legend explained below:

- *Haemophilus influenzae* **Seven** Hi subfamilies are represented by varying shades of red (HAEIN_SF1, HAEIN_SF2…)
- *Haemophilus haemolyticus*
- *Neisseria mucosa*
- *Kingella denitrificans*
- *Legionella pneumophila*
- *Kingella oralis*
- *Helicobacter pylori*



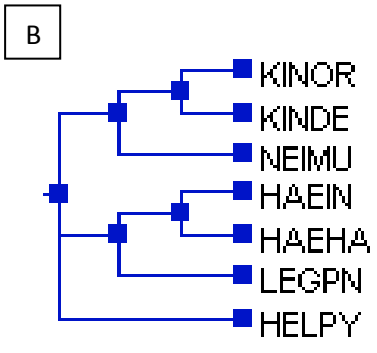


Figure 5. Species trees obtained from NCBI. Tree A was modified into Tree B before tree reconciliation.

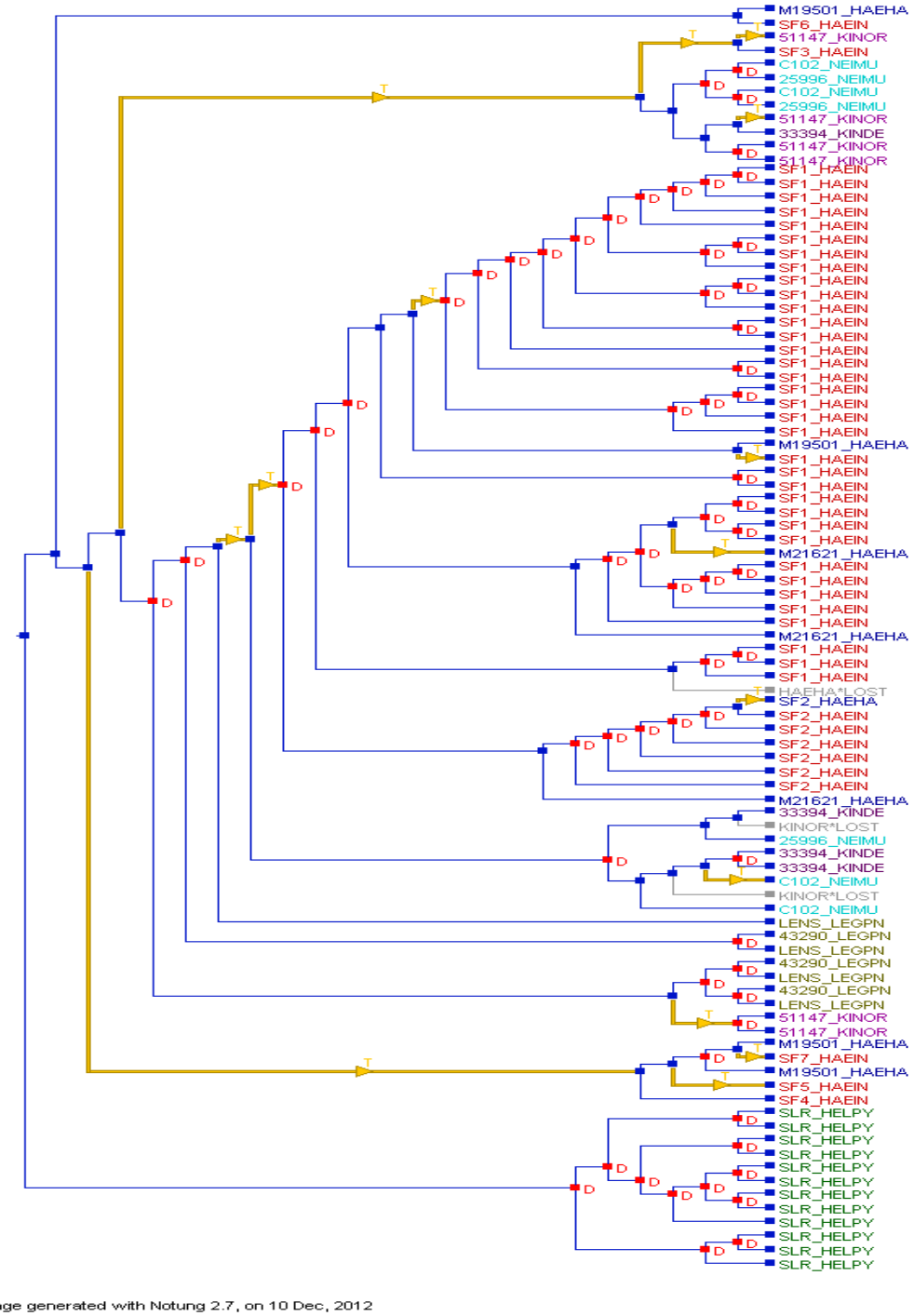


Figure 6. Reconciled SLR tree obtained from Notung 2.7. Transfers are depicted in yellow branches (T)

Table 1. **Data set**. Bacterial species shortlisted for the SLR phylogeny estimation. Corresponding columns show the number of different strains selected based on BLAST searches to representative sequence of each of the seven NTHi subfamilies of SLR.

Species	**No. of strains used**	**No. of sequences**
Haemophilus influenzae	24	47
Haemophilus haemolyticus	2	7
Helicobacter pylori	1	12
Neisseria mucosa	2	6
Legionella pneumophila	2	10
Kingella denitrificans	1	5
Kingella oralis	1	6

REFERENCES

1. Baumdicker F, Hess WR, Pfaffelhuber (2010) The diversity of a distributed genome in bacterial populations. *The Annals of Applied Probability*. 20 (2):1567-1606

2. Ehrlich GD, Ahmed A, Earl J, Hiller NL, Costerton JW, Stoodley P, Post JC, DeMeo P, Hu FZ (2010) The distributed genome hypothesis as a rubric for understanding evolution *in situ* during chronic bacterial biofilm infectious processes. FEMS Immunol Med Microbiol. 59(3):269-79

3. Boskey ER, Telsch KM, Whaley KJ, Moench TR, Cone RA (1999) Acid production by vaginal flora in vitro is consistent with the rate and extent of vaginal acidification. Infect. Immun. 67:5170-175

4. Rosenberg E, Koren O, Reshef L, Efrony R, Zilber-Rosenberg (2007) The role of microorganisms in coral health, diseases and evolution. Nat. Rev. Microbiol, 355-362

5. Goodacre R (2007) Metabolics of a superorganism. J Nutr. 259S-266S

6. Murphy TF. (2003) Respiratory infections caused by non-typeable Haemophilus influenzae. Curr Opin Infect Dis.16(2).

7. Sunakawa K, Takeuchi Y, Iwata S. (2011) Nontypeable Haemophilus influenzae (NTHi) epidemiology. Kansenshogaku Zasshi. 85(3):227-37

8. Bennett JK, Hiller NL, Janto B, Eutsey R, Powell E, Longwell M, Blackwell T, Byers B, Post CJ, Hu FZ, Ehrlich GD. Identification and Characterization of a Novel Bacterial Virulence Factorin *Haemophilus influenzae* (in preparation).

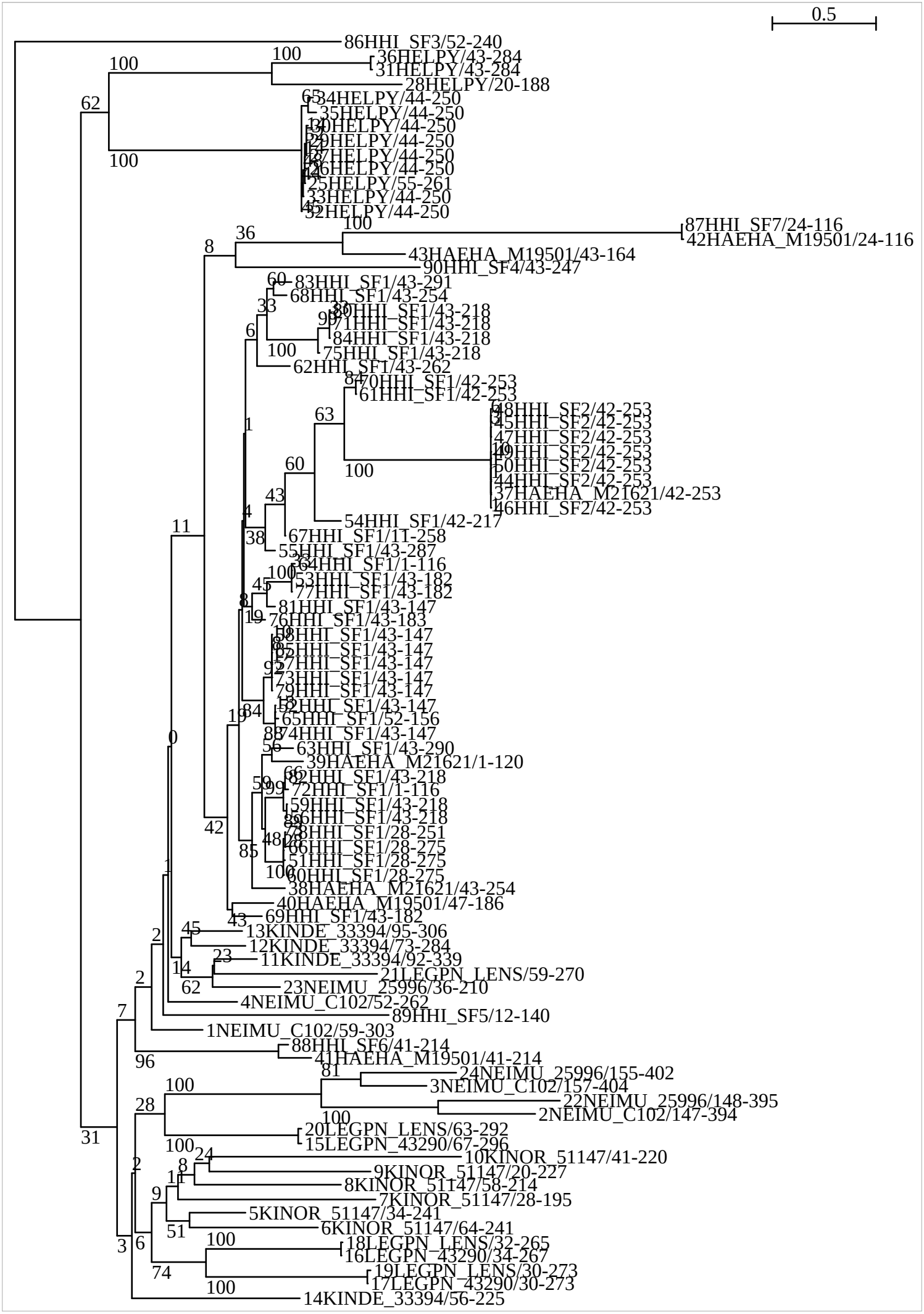
9. Newton HJ, Sansom FM, Dao J, McAlister AD, Sloan J, Cianciotto NP Hartland EL (2007) Sel1 repeat protein LpnE is a Legionella pneumophila virulence determinant that influences vacuolar trafficking. Infect Immun 75: 5575-85.

10. Ogura M, Perez JC, Mittl PR, Lee HK, Dailide G, Tan S, Ito Y, Secka O, Dailidiene D, Putty K, Berg DE, Kalia A (2007) Helicobacter pylori evolution: lineage-specific adaptations in homologs of eukaryotic Sel1-like genes. PLoS Comput Biol 3: e151.

11. Dumrese C, Slomianka L, Ziegle, U, Choi SS, Kalia A, Fulurija A, Lu W, Berg DE, Benghezal M, Marshall B, Mittl PR (2009) The secreted Helicobacter cysteine-rich protein A causes adherence of human monocytes and differentiation into a macrophage-like phenotype. FEBS Lett 583: 1637-43

12. Murphy TF, Brauer AL, Sethi S, Kilian M, Cai X, Lesse AJ. (2007) Haemophilus haemolyticus: a human respiratory tract commensal to be distinguished from Haemophilus influenzae. J. Infect. Dis. 195:81–89.

13. McCrea KW, Xie J, LaCross N, Patel M, Mukundan D, Murphy TF, Marrs CF, Gilsdorf JR. (2008) Relationships of nontypeable Haemophilus influenzae strains to hemolytic and nonhemolytic Haemophilus haemolyticus strains. J. Clin. Microbiol. 46:406–416.


0.5
86HHI_SF3/52-240
36HELPY/43-284
31HELPY/43-284
28HELPY/20-188
34HELPY/44-250
35HELPY/44-250
30HELPY/44-250
29HELPY/44-250
27HELPY/44-250
26HELPY/44-250
25HELPY/55-261
33HELPY/44-250
32HELPY/44-250
87HHI_SF7/24-116
42HAEHA_M19501/24-116
43HAEHA_M19501/43-164
90HHI_SF4/43-247
83HHI_SF1/43-291
68HHI_SF1/43-254
80HHI_SF1/43-218
71HHI_SF1/43-218
84HHI_SF1/43-218
75HHI_SF1/43-218
62HHI_SF1/43-262
70HHI_SF1/42-253
61HHI_SF1/42-253
48HHI_SF2/42-253
45HHI_SF2/42-253
47HHI_SF2/42-253
49HHI_SF2/42-253
50HHI_SF2/42-253
44HHI_SF2/42-253
37HAEHA_M21621/42-253
46HHI_SF2/42-253
54HHI_SF1/42-217
67HHI_SF1/11-258
55HHI_SF1/43-287
64HHI_SF1/1-116
53HHI_SF1/43-182
77HHI_SF1/43-182
81HHI_SF1/43-147
76HHI_SF1/43-183
58HHI_SF1/43-147
85HHI_SF1/43-147
57HHI_SF1/43-147
73HHI_SF1/43-147
79HHI_SF1/43-147
52HHI_SF1/43-147
65HHI_SF1/52-156
74HHI_SF1/43-147
63HHI_SF1/43-290
39HAEHA_M21621/1-120
82HHI_SF1/43-218
72HHI_SF1/1-116
59HHI_SF1/43-218
86HHI_SF1/43-218
78HHI_SF1/28-251
66HHI_SF1/28-275
51HHI_SF1/28-275
60HHI_SF1/28-275
38HAEHA_M21621/43-254
40HAEHA_M19501/47-186
69HHI_SF1/43-182
13KINDE_33394/95-306
12KINDE_33394/73-284
11KINDE_33394/92-339
21LEGPN_LENS/59-270
23NEIMU_25996/36-210
4NEIMU_C102/52-262
89HHI_SF5/12-140
1NEIMU_C102/59-303
88HHI_SF6/41-214
41HAEHA_M19501/41-214
24NEIMU_25996/155-402
3NEIMU_C102/157-404
22NEIMU_25996/148-395
2NEIMU_C102/147-394
20LEGPN_LENS/63-292
15LEGPN_43290/67-296
10KINOR_51147/41-220
9KINOR_51147/20-227
8KINOR_51147/58-214
7KINOR_51147/28-195
5KINOR_51147/34-241
6KINOR_51147/64-241
18LEGPN_LENS/32-265
16LEGPN_43290/34-267
19LEGPN_LENS/30-273
17LEGPN_43290/30-273
14KINDE_33394/56-225